

UNIVERSITY OF TECHNOLOGY SYDNEY

DOCTORAL THESIS

Recognising and Describing Human Activities in a Still Image

Author:

Zheng Zhou

Supervisor:

Xiangjian He

*A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy*

in the

School of Computing and Communications
Faculty of Engineering and Information Technology

September 2017

Declaration of Authorship

This thesis is the result of a research candidate conducted jointly with another University as part of a collaborative Doctoral degree. I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as part of the collaborative doctoral degree and/or fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signed:

Date:

Abstract

Understanding human activities in a still image is an essential research branch of artificial intelligence. For a computer system, the ability of understanding activities in an image is composed of not only the ability of recognising the activities in the image but also the ability of describing the recognised activities. In the age of big data, activity recognition and description generation for image have received increasing research attention, since they are of great importance in image-based information retrieval, automated image collection and collation, human-computer interaction and automated security surveillance. This thesis conducts research on recognising and describing activities in a still image and achieves several innovative achievements as follows.

(1) A framework for recognising human activities based on analysing the interactions among people is proposed. The interactions among people provide useful context for activity recognition but have not been fully taken advantage of by the existing approaches for both individual and group activity recognition. The framework is constructed based on analysing the mechanism that human brains analyse the interactions, and composed of four key sub-tasks, including Human Detection and Segmentation, Feature Extraction, Interaction Analysis and Activity Recognition.

(2) An approach for recognising individual activities based on human-interaction analysis is developed. This approach uses an innovative single-level model, called the Non-hierarchical Interaction Analysis Model (NIAM), to analyse the interactions between individuals. The NIAM does not contain a level representing groups and a group discovery process, in order to avoid the errors occurred in and computation consumed for group discovery. Several innovative algorithms are proposed and compose the body of the recognition approach, including a Fusion Restricted Boltzmann Machine for fusing features of different dimensional scales, a Focal Subspace Measurement for calculating the interdependencies between people and a Global-Local Cue Integration Method for selecting and integrating the cues extracted from different people.

(3) An approach for recognising group activities based on human-interaction analysis is developed. This approach uses a new multiple-level generative model, called Mixed Group Activity Model (MGAM). Compared with the popular discriminative multiple-level models, the MGAM performs better in comprehensively analysing the information of multiple levels of activities and modeling the interactions among multiple individuals or groups. To connect the MGAM with the raw features in an image, a Body-Part-Angle (BPA) descriptor is proposed. The BPA descriptor is friendly to a generative model that the generation distribution between the model and the raw features can be easily defined and learned.

(4) A description generator for describing the human-object interaction activities in images with natural language is proposed. Compared with the sentences given by the traditional retrieval-based approaches, the sentences given by this generator are closer to what is really happening in an image. The generator is implemented based on a deep understanding framework with a 3D spatial layout analysis and a syntactic-tree-based language model.

Acknowledgements

Foremost, I wish to appreciate my principle supervisor Professor Xiangjian He for his patience, motivation, constant encouragement and help during my candidature. His guidance helped me in all the time of research. Without his support, I cannot finish my PhD study.

I would also like to thank my colleagues and the staff in the Faculty of Engineering and Information Technology (FEIT): Wenjing Jia, Qiang Wu, Min Xu, Haiying Xia, Fan Dong, Shuai Jiang, Khaled Aldebei, etc., for their invaluable help and support. I have the honour of studying and working with them in the past two years which is valuably stamped in my life.

Last but not least, I would like to thank my family for their unconditional support throughout my life. I met my wife Jianghan Ke in the period of my PhD study. She was the biggest treasure that I gained during my PhD time and will be the biggest treasure that I ever have during my whole life.

Contents

Declaration of Authorship	i
Abstract	ii
Acknowledgements	iv
Contents	v
List of Figures	viii
List of Tables	x
1 Introduction	1
1.1 Aims and Motivations	1
1.2 Related Work	5
1.2.1 Overview of Vision-Based Human Activity Recognition	6
1.2.2 Individual Activity Recognition Approaches	8
1.2.3 Group Activity Recognition Approaches	13
1.3 Thesis Organisation	18
2 Framework for Activity Recognition Based on Interaction Analysis	20
2.1 Introduction	20
2.2 Analysis for Activity Recognition Based on Interaction Analysis	21
2.3 Framework for Activity Recognition Based on Interaction Analysis	27
2.4 Conclusion	29
3 Individual Activity Recognition Based on Human Interaction Analysis	30
3.1 Introduction	30
3.2 Related Work	33
3.3 Overview of the Non-hierarchical Approach for Individual Activity Recognition	34
3.4 Human Detection and Feature Extraction	36
3.4.1 Human Detection	36
3.4.2 Feature Extraction	37
3.4.3 Feature Fusion with Fusion Restricted Boltzmann Machine	38
3.5 Interaction Analysis	45
3.5.1 Non-hierarchical Interaction Analysis Model	45

3.5.2	Focal Subspace Measurement	46
3.6	Activity Recognition	47
3.6.1	Local Cue Generation and Local Activity Recognition	47
3.6.2	Glocal-Local Cue Interaction Method	47
3.7	Experiments	50
3.7.1	Datasets and Experimental Settings	50
3.7.2	Performance Analysis and Discussions	52
3.7.3	Comparisons with the Other Approaches	59
3.8	Conclusion	62
4	Group Activity Recognition Based on Human Interaction Analysis	63
4.1	Introduction	63
4.2	Related Work	65
4.3	Overview of the Generative Approach for Group Activity Recognition . .	67
4.4	Human Detection and Feature Extraction	69
4.4.1	Human Detection	69
4.4.2	Body-Part-Angle Descriptor for Feature Extraction	70
4.5	Interaction Analysis and Activity Recognition	73
4.5.1	Mixed Group Activity Model	73
4.5.2	Activity Recognition Based on Mixed Group Activity Model . . .	78
4.6	Experiments	79
4.6.1	Datasets and Experimental Settings	79
4.6.2	Performance Analysis and Discussions	80
4.6.3	Comparisons with the Other Approaches	84
4.7	Conclusions	85
5	Description Generation Based on a Deep Understanding framework	87
5.1	Introduction	87
5.2	Related Work	89
5.3	Deep image understanding framework	91
5.3.1	Overview	91
5.3.2	Deep Hierarchical Model for HOI Activity Recognition	91
5.3.3	Natural Language Model for Sentence Generation	92
5.4	Graph-based Method for Object 3D Layout Analysis	93
5.4.1	Representation and Inference	93
5.4.2	Parameter Learning	95
5.5	Factored Three-Way Interaction Machine	95
5.5.1	Model Representation	95
5.5.2	Model Inference	96
5.5.3	Model Parameter Optimisation Learning	97
5.6	Natural Language Generation Model	100
5.6.1	Mapping of Spatial Layout and Preposition Structure	100
5.6.2	Syntactic Tree Generation and Translation	102
5.7	Experiments	105
5.7.1	Datasets and Experimental Settings	105
5.7.2	Test on Object Layout Analysis	106
5.7.3	Test on HOI Activity Recognition	108

5.7.4	Test on Description Generation	111
5.8	Conclusions	113
6	Conclusions and Future Work	115
6.1	Conclusions	115
6.2	Future Work	116

List of Figures

1.1	Examples of narrow-angle images	3
1.2	Examples of wide-angle images	3
1.3	The statistics of the publications on the image-based activity recognition .	6
1.4	The types of cues	8
2.1	Examples of human interactions	21
2.2	Examples of human interactions	22
2.3	The framework for activity recognition based on interaction analysis . . .	27
3.1	The overview of the activity recognition approach based on non-hierarchical interaction analysis	35
3.2	The employed human detection approach	37
3.3	The structure of FRBM	39
3.4	The Nonhierarchical Interaction Analysis Model	45
3.5	Examples of entropy curves of accumulation results	49
3.6	Examples of images in the Mixed Activity Dataset	51
3.7	Examples of images in the challenging Structured Group Dataset	51
3.8	The accuracies of the Global Prediction and baselines Local Prediction using different CNNs	53
3.9	The confusion matrices on MAD	54
3.10	The validation of the FRBM	54
3.11	The validation of the FSM	55
3.12	The validation of the cue-selection algorithm in the GLCIM	55
3.13	The validation of the cue-integration in the GLCIM	56
3.14	The validation of interdependency weight in the GLCIM	56
3.15	The overall accuracies using different active intervals.	56
3.16	The confusion matrices on SGD	58
3.17	The confusion matrices on CACD	62
4.1	The discriminative model proposed by Lan et al.	66
4.2	The generative model proposed by Li et al.	67
4.3	The overview of the MGAM	68
4.4	The 3D skeletons of 20 joints or 15 joints	70
4.5	The 2D skeleton of 15 joints used by the BPA descriptor	71
4.6	The graphical model of the MGAM	74
4.7	The confusion matrix of the generative approach	81
4.8	The visualisation of the results of the proposed generative approach . . .	82
4.9	The effect of standard pose number in the MAGM	83

4.10	The effect of ratios of $\alpha : \beta$ in the MAGM	83
5.1	Examples of images returned a keyword-based image-searching engine . . .	88
5.2	The deep image understanding framework	91
5.3	The deep hierarchical model	92
5.4	The directed graph for object 3D layout analysis	93
5.5	The visualisation of 3D spatial relationships between objects v_i and v_j . .	94
5.6	The graphical illustration of FTWIM's sketch	96
5.7	The statistical model is proposed to learn spatial mapping	101
5.8	The structures of the five subtrees	103
5.9	An example of syntactic tree generation	105
5.10	Examples of the joint dataset	106
5.11	The comparisons of object layout analysis	107
5.12	The comparisons on HOI activity recognition in terms of precision	109
5.13	The comparisons on HOI activity recognition in terms of Precision-Recall curves	109
5.14	The visualisation of the results of HOI activity recognition	110
5.15	Examples of image descriptions given by the proposed description gener- ator and 4 comparative methods	112

List of Tables

1.1	The numbers of the related publications in the conferences of CVPR, ICCV and ECCV (2014-2016)	7
3.1	The overall and individual category’s accuracies with 50% local prediction accuracy	54
3.2	The accuracies of the proposed approach on the SGD using the ground-truth locations	57
3.3	The accuracies of the proposed approach on the SGD using the automatically detected locations	57
3.4	The comparisons with a newly proposed hierarchical method	60
3.5	The comparisons with the existing methods on CACD	61
4.1	The overall and per-class accuracies and variances	80
4.2	The comparisons with the state-of-the-art methods	84
5.1	Examples of the mapping from the spatial layout of object nouns to the space of prepositional phrases	101
5.2	The automatic evaluation for results generated by the proposed description generator and 4 comparative methods	111
5.3	The human judgment for results generated by the proposed description generator and 4 comparative methods	113